MIEIC - FEUP
2011-2012

EXAME

Extracção de Conhecimento e Aprendizagem Computacional

João Mendes Moreira e José Luís Borges

1 Fevereiro 2012

Duração: 2 horas ***- Sem consulta***

Responda às perguntas com **LETRA LEGÍVEL**

PARTE I

1. (2.5 Valores) Explique de forma breve quais das seguintes actividades podem ser consideradas 'data mining':

a) Segmentar os clientes de uma empresa de acordo com a sua rentabilidade.

b) Monitorizar o batimento cardíaco de um paciente por forma a detectar anomalias.

c) Prever o resultado de o lançamento de um dado equilibrado.

d) Prever o valor em bolsa de uma empresa com base nos registos históricos.

e) Agrupar os clientes de uma empresa de acordo com os seus hábitos de consumo.

2. (2 Valores) Discuta a importância do tratamento de 'missing values' na tarefa de pré-processamento de dados. Refira os métodos que conhece para a resolução do problema de 'missing values', forneça uma breve descrição dos mesmos e indique quando estes devem ser utilizados.

3. (1.5 Valores) Considere o seguinte conjunto de itemsets de tamanho 2:

{10,20} {10,30} {20,30} {20,40}

Utilize o algoritmo Apriori para calcular o conjunto de itemsets candidatos de dimensão 3 cujo suporte deverá ser calculado.

4. (1.5 Valores) Explique como se calcula a confiança e o lift de uma regra de associação. Explique o interesse de cada uma dessas medidas.

5. (2.5 Valores) Nos métodos de clustering hierárquicos há diversas formas de calcular a distância entre clusters (linkage rules). Diga quais são e em que consistem.

PARTE II

6. (2 valores) Explique porque razão se divide o conjunto de dados em sub-conjuntos de treino e de teste quando a tarefa de aprendizagem é preditiva. Descreva duas formas habituais de fazer essa separação.

7. (2 valores) Quais são as vantagens e/ou desvantagens de discretizar variáveis numéricas para a construção de árvores de decisão?

8. (2 valores) Explique em que consiste e para que servem as matrizes de custo (*cost sensitive learning*). Dê um exemplo concreto em que seja pertinente a sua utilização.

9. (2 valores) Explique sucintamente como funciona o algoritmo de redes neuronais, *back propagation*. Explique para que serve a taxa de aprendizagem (*learning rate*).

10. (2 valores) Explique que características devem ter os modelos que constituem um *ensemble* (para ambos os casos: classificação e regressão).